ON THE TIME-WINDOW FULFILLMENT RATE IN INVENTORY SERVICE LEVEL AGREEMENTS

Derek Atkins
Sauder School of Business, University of British Columbia, Vancouver, BC, Canada V6T 1Z2
derek.atkins@sauder.ubc.ca, (1) 604 8229665
Liping Liang
Department of Computing and Decision Sciences, Lingnan University, 8 Castle Peak Road, Tuen Mun, Hong Kong
lipingliang@ln.edu.hk, (852) 2616 8103

ABSTRACT

Service-level agreements (SLAs) are widely observed in practice for managing suppliers' performance. We study the roles of performance measures in SLAs using an application to inventory management. We consider two inventory performance measures: immediate ready rate (1 - stockout rate) and time-window ready rate, which are defined as the long-run percentage of periods in which demands are completely filled immediately or within a pre-specified time window, respectively. We identify situations under which an immediate ready rate results in close-to-optimal supply chain efficiency and situations under which a time-window ready rate is preferred, and find the roles of the time window in the SLA.

Key words: service level agreement; delivery performance; time-window fulfillment rate.

EXTENDED ABSTRACT

With the increased outsourcing of manufacturing and services to suppliers comes a need for better contractual agreements between suppliers and buyers. One of the most widely employed contractual instruments is a type of performance-based contracts called Service Level Agreements (SLAs). A survey by Oblicore Inc. [3] in 2007 revealed that 91% of organizations use SLAs for managing suppliers, internal agreements, or external customer agreements.

SLAs are often used when the parties involved have a long-term relationship, where the transactions are not one time. Since a fixed price alone is not enough to guarantee the delivery of the required performance, positive and/or negative performance incentives are needed. For example, a penalty might be imposed when the supplier underperforms compared to some target service level. The penalty is not based on daily transactions but on performance over a period of time. Therefore, a fundamental issue in SLA design is what performance measure should be used to align a supplier's incentive with the buyer's objective.

Performance measurement has long been recognized as a central problem in principal-agent theory and has been extensively studied in the economics literature, but the performance measure considered there is generic with a simple form and not specific to any real-world application. The operations management researchers study performance measures for applied problems, but the focus is generally on the long-run expected value of a performance measure rather than its incentive effect. Thus it is of great importance to examine the role of a performance measure for specific operations management activities and its effectiveness for incentive alignment, and identify the appropriate performance measures to be used in an SLA.

To address the above research question, we consider an application of SLAs to inventory management and two common types of inventory service performance measures: immediate ready rate and time-window ready rate. Specifically, we study a single-item inventory system with a continuous-review base-stock policy, stochastic and stationary demand, and full backlogging. We consider a supply chain consisting of a single supplier and a single buyer, where the supplier can invest both in inventory and in the inventory replenishment lead time to meet a service level target, and both investments are unobservable to the buyer. The supplier owns the inventory and incurs a linear inventory holding cost. For each delayed delivery, the buyer incurs a cost which is a convex and increasing function of the amount of delay. An SLA uses a multi-period review strategy, under which the supplier's inventory performance is reviewed every R periods (called a review phase), and if it is below a pre-specified performance threshold, then the supplier will pay a penalty linear in the amount of performance deviation from the threshold.

Fill rate and stockout rate are inventory performance measures commonly used in both the practice and the literature. Most inventory management literature studies performance measures in the long run using expected performance. But the performance measure in a finite review phase is a random variable. When the supplier's actions are unobservable, it is important to know the distribution of the performance measure in order to provide an incentive to the supplier. Since it is very difficult to derive the distribution of fill rate, we focus mainly on the ready rate, which is the long-run fraction of time that demands are filled immediately from the stock. It measures inventory availability, and is equal to $1-$ stockout rate. The conventional ready rate and fill rate are measures of the off-the-shelf or immediate order fulfillment performance. In practice, time-window fulfillment rates are more commonly used than off-the-shelf performance measures (LaLonde et al. [2]). In Quick Response and other forms of time-based competition, the performance measure for customer service is often the ability to meet delivery promises, where the promised time window is usually small. For example, around 1995, Hewlett-Packard aimed at a 93% fulfillment rate within 3 days, and IBM PC and Compaq 95% within 5 days (Hausman et al. [1]). When the performance measure is based on on-time delivery by a supplier to a buyer, time-window fulfillment rates are also used. Therefore, we also study another form of ready rate, the ready rate with a window, which is the long-run fraction of time that demands are filled within a pre-specified time window. We study the ready rate for simplicity in exposition, but similar insights can be obtained when either the immediate or time-window fill rate is used as the performance measure.

The objective of this paper is twofold. First, we address the design of SLAs in supply management, including choosing the performance measure, and determining the performance target, the allowable performance deviation from the target, and the penalty for underperformance. Specifically, we examine two types of SLAs using either the immediate or time-window ready rate as the performance measure. Second, we compare these two forms of SLAs in terms of the average supply chain cost. We show that when the supplier employs a static inventory policy, can invest both in inventory level and in supply lead time, with the investments unobservable to the buyer, an SLA using the time-window ready rate can induce the supplier to make the investments compatible with overall supply chain optimization. An SLA using only the immediate ready rate generally cannot induce this first-best investment. We also discuss the issue of using a single performance measure for aligning the supplier's incentive when the supplier has multiple ways to achieve inventory performance. We find that the time window in the performance measures plays three roles: (i) it aligns the supplier's tradeoff between inventory and lead time investments with that of the supply chain; (ii) it allocates inventory risk between the buyer and a supplier; and (iii) to some extent it transfers the buyer's delay cost structure to the supplier.

References

[1] W. H. Hausman, H. L. Lee, and A. X. Zhang. Joint demand fulfillment probability in a multi-item inventory system with independent order-up-to policies. *European Journal of Operational Research*, 109(3):646–659, 1998.

[2] B. LaLonde, M. Cooper, and T. Noordwier. Customer service: A management perspective. Technical report, Council of Logistics Management, Oak Brook, IL, 1988.

[3] Oblicore Inc. 2007 service level management survey: Results, trends and analysis, 2007.